CROP BREEDING AND APPLIED BIOTECHNOLOGY

ARTICLE

Adjusting the Scott-Knott cluster analyses for unbalanced designs

Thiago Vincenzi Conrado[1*], Daniel Furtado Ferreira[1], Carlos Alberto Scapim[3] and Wilson Roberto Maluf[2]

Crop Breeding and Applied Biotechnology
17: 1-9, 2017
Brazilian Society of Plant Breeding.
Printed in Brazil
http://dx.doi.org/10.1590/1984-70332017v17n1a1

Abstract: *The Scott-Knott cluster analysis is an alternative approach to mean comparisons with high power and no subset overlapping. It is well suited for the statistical challenges in agronomy associated with testing new cultivars, crop treatments, or methods. The original Scott-Knott test was developed to be used under balanced designs; therefore, the loss of a single plot can significantly increase the rate of type I error. In order to avoid type I error inflation from missing plots, we propose an adjustment that maintains power similar to the original test while adding error protection. The proposed adjustment was validated from more than 40 million simulated experiments following the Monte Carlo method. The results indicate a minimal loss of power with a satisfactory type I error control, while keeping the features of the original procedure. A user-friendly SAS macro is provided for this analysis.*

Key words*: Type I error rate, unequal number of observations, Monte Carlo simulations, means clustering procedures, SAS macro.*

INTRODUCTION

A common problem in plant breeding is comparison of new genetic combinations. In order to detect significant difference among treatments, several Multiple Comparison Procedures (MCP) were developed: LSD (Fisher 1935), Tukey (1949), SNK (Student 1908, Newman 1939, Keuls 1952), Scheffé (1953), and Duncan (1955). Nonetheless, all these procedures can result in groups overlapping, where one treatment ends up belonging to two or more groups simultaneously (Calinski and Corsten 1985). This behavior usually prevents a clear division of the whole set into two or more groups of treatments and also leads to a more complex simultaneous analysis of multiple variables due to the presence of overlapping subsets. Thus, selection for advancement of new genetic combinations to the next step in the plant breeding program requires extra effort to overcome this statistical issue.

Cluster analysis is a promising solution to avoid subset overlapping from widely-used MCP (O'Neill and Wetherill 1971, Plackett 1971). One example of an intuitive and satisfactory approach, avoiding subset overlapping, is the use of cluster analysis over the Mahalanobis generalized distance (Rao 1952). Additionally, clustering techniques can be applied to taxonomy purposes since they have high affinity to Hotelling's Principal Component Analysis and Fisher's Discriminant Analysis (Hotelling 1933, Fisher 1936, Edwards and Cavalli-Sforza 1965).

***Corresponding author:**
E-mail: tconrado@hotmail.com

Received: 28 June 2015
Accepted: 07 July 2016

[1] Universidade Federal de Lavras (UFLA), Departamento de Biologia, Campus, CP 3037, 37.200-000, Lavras, MG, Brazil
2 UFLA, Departamento de Agricultura
3 Universidade Estadual de Maringá (UEM), Agronomia, 87.080-000, Maringá, PR, Brazil

In 1974, Alastair J. Scott and Martin Knott publicized their idea of using the Maximum Likelihood (ML) ratio test to evaluate the significance of partitions from cluster analysis of sample treatment means in designs with an equal number of observations per treatment (Scott and Knott 1974). The first review of methods for Scott-Knott means separation suggesting their use in agronomics was provided several years afterward (Chew 1976). The Scott-Knott approach is an alternative to the MCP in a situation in which two or more internally homogenous subsets of sample treatment means are expected. It uses a univariate form of the divisive clustering procedure (Edwards and Cavalli-Sforza 1965) with a likelihood ratio test for determining when to stop the clustering process to create non-overlapping, distinct, and exclusive subsets of sample treatment means. The process orders the treatment means to minimize the number of possible treatment mean partitions to be pondered (Fisher 1958) and then maximizes the sum of squares between clusters to determine the best partitioning. Despite a significant increase in the calculation volume for every additional treatment even after the ordering of treatment means, it is still feasible, even manually, if the number of partitions remains lower than 12 (Scott and Knott 1974). Indeed, the computations are more onerous than an MCP (Carmer and Walker 1985). Nevertheless, it should not be a problem for any modern computer (Gates and Bilbro 1978).

Some procedures with the same idea of partitioning means into non-overlapping groups were published after Scott-Knott (1974). These procedures presented variations in regard to the decision-making process and the clustering logic, ranging from agglomerative to divisive, hierarchical to non-hierarchical, but all of them ensure groups with no overlapping (Jolliffe 1975, Cox and Spjotvoll 1982, Calinski and Corsten 1985, Bozdogan 1986, Bautista et al. 1997, Di Renzo et al. 2002, Ciampi et al. 2008).

Many researchers prefer cluster analysis in order to facilitate interpretation and presentation of results since it results in non-overlapping, distinct, mutually exclusive groupings of the observed treatment means (Gates and Bilbro 1978, Carmer and Lin 1983, Calinski and Corsten 1985, Carmer and Walker 1985). This advantage is very clear when it is necessary to evaluate more than one variable simultaneously because the test easily allows for a positive selection of primary traits and a negative selection for any traits remaining to be evaluated. It can be effortlessly performed over the clustered data with multiple variables by initially applying filters to keep only higher performance clusters for the most important trait (i.e. yield) and then by removing some clusters of lower performance in the variables of secondary importance (i.e. plant height, biomass, etc.). This procedure should result in a highly reduced subset of treatments that present higher performance for the top priority trait, with a desirable level for the secondary traits.

An early evaluation of the Scott-Knott test with agglomerative procedures under scenarios where there is more than one true group of treatment means, or partial true null hypothesis (p-H_0), exposed the lack of an appropriate experimentwise type I error control. The result of simulations suggested that the test should be used only when the experiment has been performed with great precision, and it may be unsuitable for experiments where use of MCP would be considered inappropriate, such as those whose design and purpose suggest meaningful, orthogonal, linear contrasts with a single degree of freedom among the treatment means. However, the Scott-Knott test exhibited a higher ability to correctly reject the null hypothesis (power) and detect small differences between treatments than even the LSD test (Willavise et al. 1980).

Moreover, the Scott-Knott test has the highest rate of correct decisions and aptitude for improving performance as the number of treatments increases, in comparison with the SNK, Duncan, t-student, and Tukey tests (Silva et al. 1999, Borges and Ferreira 2003). The test exhibits higher than nominal type I error rate when evaluated in simulated scenarios in which the null hypothesis (H_0) is false for some treatments (p-H_0), although for scenarios where the null hypothesis is true for all treatments, the empirical type I error rate is under nominal levels even for the experimentwise type I error rate (Di Renzo et al. 2002, Borges and Ferreira 2003).

The Scott-Knott test also provides higher robustness compared to the MCP tests for mean separation in non-Gaussian distributions (Borges and Ferreira 2003). Despite the lack of control of type I error, the test demonstrates much higher Power than any MCP, although these two features, high robustness and power, are very common to most cluster analyses (Bautista et al. 1997, Silva et al. 1999, Di Renzo et al. 2002, Borges and Ferreira 2003). The Scott-Knott test displays similar type I and type II error in comparison to Bautista et al. (1997) and Di Renzo et al. (2002). However, its performance is superior to that of Jollieffe (1975) (Di Renzo et al. 2002).

Group homogeneity can be improved by changing the clustering approach from divisive to non-grouped treatment

clustering (Bhering et al. 2008). It usually reduces the number of significantly different clusters - slightly increasing the number of treatments grouped in each one of the different clusters. In spite of this drawback, this consequence can be useful in plant breeding scenarios in which positive selection followed by retesting is applied, since it can shift a small number of treatments from an inferior cluster to a superior one.

Since most plant breeding designs are unbalanced, the objective of this research is to adjust and validate the Scott-Knott test in order to allow its use in experiments under partially balanced incomplete block designs or balanced designs with missing plots, since the non-adjusted procedure is only applicable to balanced designs. This paper proposes a novel solution for use of the Scott-Knott test under unbalanced designs followed by its validation. In order to ease its use, a user-friendly macro for the SAS/STAT® software is also provided.

MATERIAL AND METHODS

Description of the proposed adjustment procedure

The original Scott-Knott (1974) test begins by ranking all the k treatment means to be grouped and then by calculating B_0 from the k treatments partitioned in two smaller subsets. The B_0 value is calculated for every $k-1$ possible partition, and the partition with the highest value of B_0 is tested using λ as two distinct subsets of treatment means. The test uses the circumference constant π (=3.14159...) and related adjusts to approximate the λ distribution to the χ^2 distribution. If the chi-square test with $\left(\frac{k}{\pi-2}\right)$ degrees of freedom rejects the null hypothesis, the process repeats; each one of these distinct subsets is, in turn, further subdivided until each of the final clusters is shown to be homogeneous by a likelihood ratio test on λ.

$$\lambda = \frac{\pi}{2(\pi-2)} B_0 \Big/ \hat{\sigma}_0^2 \qquad \text{(i)}$$

The statistic λ (i) depends on B_0, which is the maximum value from the sum of squares of all the possible partitions of k treatments into two groups, and on $\hat{\sigma}_0^2$, which is the maximum likelihood estimator of σ^2 for treatments under the null hypothesis.

Equation (ii) shows how υs^2 is used where s^2 represents an unbiased estimator of σ^2 associated with υ degrees of freedom, y_i is the treatment mean i, and $\bar{y}$ is the mean of all k treatments. The variable n is the number of replications, or the total number of blocks according to the experimental design.

$$\hat{\sigma}_0^2 = \left[\sum_{i=1}^{k}(y_i-\bar{y})^2 + \upsilon s^2\right] \Big/ (k+\upsilon)\,;\; s^2 = \frac{MSE}{n} \qquad \text{(ii)}$$

Since the full Means Square Error (*MSE*) model is a good measure of variance, it is used as a satisfactory term for estimation of s^2.

Equation (iii) shows the relation between the unbiased estimator s^2 and the Standard Error of the Mean SE_y, where *RMSE* is Root Mean Square Error. It is valid only under an equal number of observations for every treatment ($n_1 = n_2 = ... = n_k$). Additionally, under a balanced experimental design, SE_y has the very same value for every treatment and leads to equation (iv), the base of the proposed adjustment, where the mean of the sum of the squares of SE_y estimates s^2.

$$s^2 = \frac{MSE}{n} = \left(\sqrt{\frac{MSE}{n}}\right)^2, \text{ and } SE_{\bar{y}} = \frac{RMSE}{\sqrt{n}} = \sqrt{\frac{MSE}{n}}, \text{ thus } s^2 = \left(SE_{\bar{y}}\right)^2 \qquad \text{(iii)}$$

$$s^2 = \left(SE_{\bar{y}}\right)^2, \text{hence } s^2 = \frac{1}{k}\sum_{i=1}^{k}\left(SE_{\bar{y}_i}\right)^2 \qquad \text{(iv)}$$

Moreover, equation (iv) used in a balanced experimental design can be modified and expressed as equation (v), where a different number of observations for every treatment is also permitted. After the modification, the corrected unbiased estimator of s_c^2 can change according to the $SE_{\bar{y}i}$ of treatments in the partitioned set. Thus, in order to accommodate subsets of treatments with unequal and equal numbers of observations, s_c^2 should be calculated for every null hypothesis before testing the statistic λ against a χ^2 distribution with the associated υ degrees of freedom. Hence, for every clustering step, s_c^2 can change to adapt to the number of observation of each treatment in the current clustering process.

$$s_c^2 = \frac{1}{k}\sum_{i=1}^{k}\left(SE_{\bar{y}i}\right)^2 \quad \text{(v)}$$

Along with correction of s_c^2, the raw treatment mean y_i should be replaced by $\hat{y}_i$, which is the treatment mean adjusted to the effect of the unequal number of replications/blocks. The following changes in the original procedure are minimal and are disclosed in equations (vi). The notation λ_c should be used to identify λ statistics while using the correction even though the testing process against the χ^2 distribution remains the same as the original procedure.

$$\lambda_c = \frac{\pi}{2(\pi-2)}B_0 \Big/ \hat{\sigma}_{0c}^2 \text{, where } \hat{\sigma}_{0c}^2 = \left[\sum_{i=1}^{k}(\hat{y}_i - \bar{y})^2 + vs_c^2\right] \Big/ (k+v); \quad \text{(vi)}$$

As expected, the correction increases the $\hat{\sigma}_{0c}^2$ value as the number of observations per treatment decreases - lowering the final λ_c value. This leads to a lower probability of rejecting the null hypothesis, which protects the test from the type I error. The unbalanced treatment adjustment maintains the same features and results as the original method in balanced treatment scenarios. Indeed, s_c^2 only changes for clusters in an unbalanced condition (*i.e.*, missing plots). When clustering the same experiment, after partitioning all treatment means with missing plots, the remaining clusters should have the same s_c^2 value. It is important to keep in mind that since the process follows a hierarchical clustering sequence, the very same subset of treatment means with an unequal number of observations can be partitioned multiple times before composing the final specific cluster. Indeed, the calculation of s_c^2 for every candidate partition that challenges the χ^2 distribution makes the adjustment hard to be calculated manually, but it provides satisfactory protection to the original Scott-Knott test without a significant reduction in power.

Validation of the proposed adjustment procedure

The s_c^2 deduction can indicate how the correction affects the Scott-Knott test; nevertheless, it is necessary to quantify and compare the power and type I error of the adjustment while using it. in order to validate the proposed adjustment, use of the Monte Carlo method (Metropolis and Ulan 1949) is a suitable option to simulate experiments with known parameters and then evaluate the results by comparing the original test against the adjusted solution for unbalanced designs (Carmer and Swanson 1971, Silva et al. 1999, Borges and Ferreira 2003). For that purpose, more than 40 million experiments were simulated for multiple unbalanced levels combined with several α values. The simulation scheme is composed of three main branches: complete H_0 ($\mu_1 = \mu_2 = \mu_3 = \ldots \mu_l$), partial H_0 ($\mu_1 = \ldots = \mu_{l/2} \neq \mu_{(l/2+1)} = \ldots \mu_l$), and complete alternative hypothesis H_1 ($\mu_1 \neq \mu_2 \neq \mu_3 \neq \ldots \mu_l$). The first branch was used only to quantify type I error and the third only to measure power, while the second branch measures type I error and power.

All three branches contained nine levels of α (0.01, 0.02, 0.05, 0.08, 0.10, 0.12, 0.15, 0.18, and 0.20). Within each α level, there were ten levels of missing data (0.00, 0.01, 0.02, 0.05, 0.08, 0.10, 0.12, 0.15, 0.18, and 0.20). Since the second and third branches were used to evaluate the test Power they also exhibited four (1, 2, 3, and 4) levels of δ (true difference between two treatment means). In order to improve the robustness of the study, 50,000 experiments were simulated for all 810 Monte Carlo simulation setups across all three branches, culminating in a total of 40.5 million simulated experiments.

Furthermore, every simulated experiment was composed of a random number of blocks (3 to 20) and a random number of treatments (4 to 100). Experiments with a number of observations lower than 50 were replaced to avoid a small number of degrees of freedom after data removal at random to reach the required missing level. The number of both blocks and treatments were from a uniform distribution. The effects of block and observation error were from a

normal distribution with a mean of zero and a standard deviation of one. The differences between subsets were defined as the product of δ and $\sigma_{\bar{x}}$ (standard error of the mean). After each experiment was generated, some plot values were removed at random. As the simulation removed plots randomly with no restriction, the minimum number of plots was set to one per treatment to avoid treatments with no plots.

Instead of measuring type I error per comparison, it was measured per experiment, a situation in which rejection of a single incorrect null hypothesis in an experiment scores as experimentwise type I error. This approach is more severe and general because it does not consider the number of treatments in the experiment (*i.e.*, a higher number of treatments promotes an even higher number of contrasts, and it implies a higher probability of type I error). However, this approach should be able to make a better distinction between the original and adjusted procedures. Converging results were expected for both procedures (original and adjusted) under balanced designs. Thus, contrast can be observed only between balanced and unbalanced designs.

All 40.5 million experiments were simulated in SAS/IML® and analyzed with SAS System for Windows 9.3 (SAS Institute 2011). The data were evaluated using the Generalized Linear Models Procedure (Proc GLM). Output of the adjusted means was grouped by a compiled macro. A recursive SAS local host multithread approach with isolated workplaces was used to speed up the simulation run time.

Stability of the process and the ability to suspend it was ensured by the use of macros capable of error handling, also oriented to processing batches of 5,000 experiments and logging all the processing responses.

Regarding the accuracy of the estimated type I error rates using Monte Carlo simulations, the exact binomial test was applied, contrasting the nominal significance level against the obtained empirical rate (Leemis and Trivedi 1996). In scenarios in which the exact binomial test rejected the null hypothesis ($p < 0.01$), the performance of the Scott-Knott test should be considered *conservative* when the empirical rate is lower than the nominal rate and should be considered *liberal* if higher. In scenarios in which the exact binomial test did not reject the null hypothesis, the tests were classified as *precise*. The F-value was obtained using equation (vii), where y represents the number of experiments with at least one type I error, α is the nominal significance level, and N is the number of simulated experiments (50,000). The *p-value* was found using $\upsilon_1 = 2(N - y)$ and $\upsilon_2 = 2(y + 1)$ degrees of freedom (Santos et al. 2001).

$$F = \left(\frac{y+1}{N-y}\right)\left(\frac{1-\alpha}{\alpha}\right) \quad \text{(vii)}$$

RESULTS AND DISCUSSION

Table 1 summarizes the results of 4.5 million simulated experiments. These experiments were simulated under the complete H_0 hypothesis (no real difference among treatments). For experiments with a balanced design (no missing plots), as the nominal α level increased, the empirical experimentwise type I error became higher. This persisted under experiments with missing plots using the proposed Scott-Knott adjustment, but reduced when the level of imbalance increased. It can be observed that the empirical values obtained using the Monte Carlo method for a balanced design (0%

Table 1. Empirical experimentwise type I error under no true difference between treatments

Nominal Alpha	Unbalance level (%)									
	0	1	2	5	8	10	12	15	18	20
1	0.932	0.926	0.834†	0.820†	0.896	0.760†	0.776†	0.760†	0.778†	0.672†
2	1.910	1.920	1.768	1.758	1.746†	1.728†	1.724†	1.736†	1.554†	1.692†
5	4.854	4.762	4.918	4.914	4.804	4.524†	4.358†	4.558†	4.318†	4.316†
8	8.046	8.168	7.832	7.760	7.686†	7.596†	7.556†	7.356†	7.190†	7.106†
10	10.184	10.334	10.284	9.830	9.936	9.546†	9.634†	9.498†	9.500†	9.514†
12	12.436†	12.374	12.166	12.024	12.018	11.728	11.814	11.576†	11.192†	11.234†
15	15.366	15.728†	15.430†	15.248	15.052	15.058	15.062	14.602	14.580†	14.060†
18	18.686†	18.910†	18.394	18.446†	18.284	18.120	18.200	17.658	17.750	17.382†
20	20.982†	20.900†	20.614†	20.508†	20.370	20.444	19.878	19.706	19.800	19.840

† represents scenarios where the exact binomial test rejected the null hypothesis

of missing plots), in which the adjusted and non-adjusted procedures exhibit the same results, are below the nominal α level for values smaller than 0.05, but according to the exact binomial test, the difference is not significant. In contrast, the empirical value is significantly higher than the nominal value for some α levels higher than 0.10, which means that the original procedure should be considered liberal at these levels since it does not properly control the type I error even under the complete H_0 hypothesis. The intermittent classification for the alpha levels of 0.12 and 0.15 as a trend for empirical rates to surpass nominal rates as the nominal alpha level increases could be caused by approximation to the χ^2 distribution used by Scott-Knott (1974), but this thesis should be evaluated in further studies and does not belong to the scope of this study.

Moreover, in half of the simulated combinations, the experimentwise type I error was evaluated as significantly different from the nominal value by the exact binomial test. As expected, the adjustment led to a more conservative approach as the level of missing plots increased. This result suggested that in order to use the proposed adjustment, the user must take into account the level of imbalance (either from the planned design or from random loss of plots) before selecting the nominal α level.

In contrast, the adjusted and non-adjusted (original) Scott-Knott test exhibited a higher empirical experimentwise type I error rate than the nominal rate under p-H_0 (Table 2). It also showed a small increase in the experimentwise type I error rate when the level of missing plots became higher, but the magnitude of the experimentwise type I error rate reduced as the α level increased. This result validated the findings of Silva et al. (1999) and exposed the weakest point of the Scott-Knott test - the lack of control of experimentwise type I error under a p-H_0.

Additionally, lower values of δ culminated in smaller differences in the experimentwise type I error rate between the adjusted and non-adjusted results of the Scott-Knott procedure (Figure 1). This trend persisted upon increasing the nominal α. Increasing α or δ led to a reduction in the difference in Power among balanced and unbalanced experimental designs (Table 3). It is also important to keep in mind that a higher value of δ indicates larger differences among the treatment values. Hence, it is easier for both procedures to detect these differences and reject the null hypothesis for any level of imbalance. The adjusted and non-adjusted tests exhibited lower Power for $\delta \leq 1$. No significant differences in Power between the adjusted and non-adjusted procedures were noticed for $\delta > 1$. Additionally, the adjusted Scott-Knott test maintained very high Power, even with a small α value under a complete H_1 (Figure 2).

However, as the level of imbalance got higher, there was

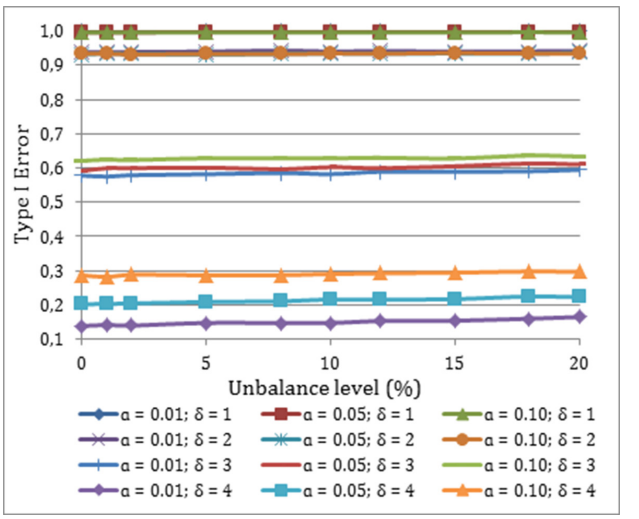

Figure 1. Empirical experimentwise error under the partial null hypothesis in the combination of three significance levels (α) by four levels of true difference between two treatment means (δ).

Table 2. Empirical experimentwise type I error under true difference between treatments of four standard errors of the mean ($4\sigma_{\bar{x}}$)

Nominal Alpha	Unbalance level (%)									
	0	1	2	5	8	10	12	15	18	20
1	13.842	14.136	13.962	14.748	14.722	14.740	15.280	15.398	15.986	16.482
2	15.124	15.560	15.740	15.870	16.474	16.504	16.860	17.132	17.374	17.894
5	20.218	20.280	20.456	20.830	21.100	21.558	21.532	21.692	22.472	22.246
8	25.406	25.408	25.136	25.244	25.944	25.798	25.974	26.114	26.246	26.952
10	28.676	28.178	28.818	28.674	28.706	29.046	29.222	29.440	29.756	29.684
12	31.684	31.522	31.628	31.670	31.874	31.722	32.100	31.938	32.492	32.448
15	36.538	36.356	36.696	36.192	36.470	36.186	36.368	36.632	36.688	36.554
18	40.600	40.778	40.530	40.698	40.960	40.770	40.602	40.984	41.238	41.174
20	43.680	43.630	43.438	43.448	43.514	43.486	43.530	43.846	43.260	43.600

Table 3. Power of Adjusted Scott-Knott in several unbalance levels under the partial null hypothesis (H_0) under four levels of true difference between two treatment means (δ)

δ	Unbalance level (%)									
	0	1	2	5	8	10	12	15	18	20
p=0.01										
1	32.525	32.652	32.233	31.884	31.453	30.982	31.306	30.748	30.236	29.735
2	84.938	84.993	85.082	85.029	85.062	85.107	85.015	85.014	85.067	85.027
3	96.582	96.574	96.566	96.537	96.516	96.513	96.491	96.459	96.452	96.433
4	99.519	99.513	99.515	99.484	99.469	99.477	99.454	99.438	99.415	99.397
p=0.05										
1	48.049	47.38	47.468	47.305	46.570	46.849	46.455	46.399	45.756	45.600
2	85.206	85.256	85.253	85.295	85.259	85.289	85.289	85.292	85.339	85.228
3	96.662	96.673	96.605	96.627	96.638	96.625	96.596	96.565	96.529	96.534
4	99.552	99.546	99.538	99.514	99.500	99.475	99.476	99.456	99.430	99.430
p=0.10										
1	53.764	53.616	53.668	53.400	53.436	52.877	53.281	52.663	52.678	52.265
2	85.406	85.338	85.369	85.365	85.362	85.396	85.386	85.453	85.439	85.464
3	96.792	96.786	96.794	96.757	96.710	96.718	96.694	96.690	96.659	96.653
4	99.560	99.559	99.542	99.542	99.532	99.509	99.502	99.479	99.459	99.459

a small loss of power when using the proposed adjustment. This performance was expected since missing information causes lower ability to reject the null hypothesis due to the additional protection required to control type I error. The small loss of power is a suitable indicator for adjustment efficiency, which is very important since the Scott-Knott test is recognized for its high power, with superior performance over the LSD and other widely used MCP (Willavise et al. 1980, Silva et al. 1999, Borges and Ferreira 2003). In spite of that, there is a trend of power reduction as the number of members per cluster decreases. This has already been pointed out and is similarities between hierarchical and non-hierarchical procedures (Tasaki et al. 1987), but it should not be assumed to be common to all clustering procedures since the clustering procedure of Bozdogan (1986) shows exactly the opposite response.

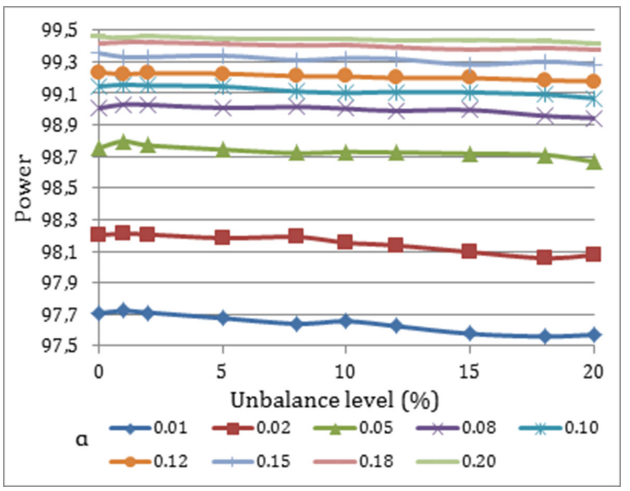

Figure 2. Empirical power under the complete H_1 hypothesis in nine significance levels (α) across ten unbalance levels.

Although the loss of power lowers the total number of clusters, it is a tolerable deficiency for scenarios where the entries that are wrongly clustered together should be retested in the next stage of research. Since the retesting routine is often used in plant breeding programs, this error is preferable to the possibility of the error of discarding an entry without a satisfactory level of confidence. Thus, as for the non-adjusted Scott-Knott procedure, it is necessary to understand the error tolerance of the experiment under evaluation before using the proposed adjustment.

It is noteworthy that even using the proposed adjustment, the most common cause of the type I error under p-H_0 for the Scott-Knott test is late compensation for incorrect partitioning in the previous step, as a consequence of divisive binary partitioning. This usually occurs in scenarios where the true number of clusters is different from powers of 2 or from the geometric sequences with common ratio 2 (data not shown). This unsatisfactory compensation is very noticeable when the true number of clusters is 3, which is a weakness common to various clustering procedures (Tasaki et al. 1987). If the gap between clusters is not clear enough, the maximum likelihood test may select a splitting point around the median by mistake. Then, in the next step, while it seeks for the point that maximizes the likelihood, it has a chance to correctly split the subset between the first and second clusters. A clear demonstration of this is an experiment with 9 treatments

truly distributed in 3 clusters, for example ABC/DEF/GHI, in which the test incorrectly performs the first partitioning as ABCDE/FGHI and then it differentiates the first true cluster from the rest of the subset, resulting in (ABC/DE)/FGHI. In following, for the same reason, the test can correctly discriminate the third true cluster from treatment F, culminating in 4 clusters: (ABC/DE)/[F/GHI]. Although the first and third clusters are correct, the second cluster is improperly divided, increasing the type I error rate. This type of result is a consequence of adoption of a divisive hierarchical approach in order to allow comparison of the selected critical value which was obtained by empirical approximation, and afterwards, to declare the computed statistic significant or not (Carmer and Lin 1983). Some approaches avoiding hierarchical clustering have been published to avert this undesirable feature by simply allowing the creation of completely new clusters in every step of evaluation (Cox and Spjotvoll 1982, Calinsk and Corsten 1985, Bozdogan 1986). Despite that, the divisive hierarchical approach is still used for clustering (Di Rienzo et al. 2002, Valdano and Di Rienzo 2007).

Within plant breeding applications, the use of non-overlapping, mutually-exclusive subsets such as Scott-Knott creates a clear cutoff for the genotype advancement procedure, while results with multiple distinct subsets can help in financial management by assigning the right subset to an appropriate testing pipeline. Using the proposed adjustment procedure, this distinguishing feature is extended to experiments with missing data, which are very common in yield trials. For example, using cluster analysis on an unbalanced yield trial that results in 6 distinct subsets, the breeder would be able to submit solely the genotype subset partitioned in the highest category, “Group A”, to be tested in the most accurate and expensive Pipeline I (the maximum number of locations in a randomized complete block design). Group B of genotypes could be placed in the intermediate Pipeline II (a smaller set of locations), and Group C and D could be tested in the lower cost Pipeline III (augmented blocks in the same locations as Pipeline II), while discarding the genotypes in Groups E and F (that have inferior performance compared to the commercial checks, clustered in Group C). After harvesting, the breeder can choose to retest only the superior genotypes from Pipeline III together with the new entries to be tested in Pipeline II or I.

A small drawback to the use of the proposed adjustment procedure is the increased complexity and volume of calculations in comparison to the non-adjusted procedure. Thus, in order to promote better dissemination of the proposed adjustment, a free compiled SAS GLM macro was developed and can be downloaded at http://www.tconrado.com/sas/sk.zip. The compressed file also contains an example to provide better understanding of the macro options and about how to use the software.

The proposed adjusted Scott-Knott procedure had performance similar to the original procedure under unbalanced experimental designs, with minimal loss of power, while maintaining satisfactory control of the experimentwise type I error and improved performance at $\alpha \geq 0.05$. This adjustment increases the spectrum for use of the test, providing the researcher with an alternative to the MCP, even under a significant loss of experimental data (missing plots), and it is readily available for use in SAS.

ACKNOWLEDGMENTS

We appreciate the helpful comments made by Mr. Gregory Reeves and the support from the Brazilian Government offered by the National Council for Scientific and Technological Development (CNPq) and the National Council for the Improvement of Higher Education (CAPES).

REFERENCES

Bautista MG, Smith DW and Steiner RL (1997) A cluster-based approach to means separation. **Journal of Agricultural, Biological and Environmental Statistics 2**: 179-197.

Bhering L, Cruz CD, Vasconcelos ES, Ferreira A and Resende MFR (2008) Alternative methodology for Scott-Knott test. **Crop Breeding and Applied Technology 8**: 9-16.

Borges LC and Ferreira DF (2003) Poder e taxas de erro tipo I dos testes Scott-Knott, Tukey e Student-Newman-Keuls sob distribuições normal e não normais dos resíduos. **Revista de Matemática e Estatística 21**: 67-83.

Bozdogan H (1986) Multi-sample cluster analysis as an alternative to multiple comparison procedures. **Bulletin of Informatics and Cybernetics 22**: 95-130.

Calinski T and Corsten LCA (1985) Clustering means in ANOVA by simultaneous testing. **Biometrics 41**: 39-48.

Carmer SG and Lin WT (1983) Type I error rates for divisive clustering methods for grouping means in analysis of variance **Communications in Statistics - Simulation and Computation 12**: 451-466.

Carmer SG and Swanson MR (1971) Detection of differences between

means: a Monte Carlo study of five pairwise multiple comparison procedures. **Agronomy Journal 63**: 940-945.

Carmer SG and Walker WM (1985) Pairwise multiple comparisons of treatment means in agronomic research. **Journal of Agronomic Education 14**: 19-26.

Chew V (1976) Comparing treatment means: a compendium. **Hortscience 11**: 348-357.

Ciampi A, Lechevallier Y, Limas MC and Marcos AC (2008) Hierarchical clustering of subpopulations with a dissimilarity based on the likelihood ratio statistic: application to clustering massive data sets. **Pattern Analysis and Applications 11**: 199-220.

Cox DR and Spjotvoll E (1982) On partitioning means into groups. **Scandinavian Journal of Statistics 9**: 147-152.

Di Rienzo JA, Guzmán AW and Casanoves F (2002) A multiple-comparisons method based on the distribution of the root node distance of a binary tree. **Journal of Agricultural, Biological, and Environmental Statistics 7**: 129-142.

Duncan DB (1955) Multiple range and multiple F tests. **Biometrics 11**: 1-42.

Edwards AWF and Cavalli-Sforza LL (1965) A method for cluster analysis. **Biometrics 21**: 362-375.

Fisher RA (1935) **The design of experiments**. Oliver and Boyd, London, 252p.

Fisher RA (1936) The use of multiple measurements in taxonomic problem. **Annals of Eugenics 7:** 179-188.

Fisher RA (1958) On grouping for maximum homogeneity. **Journal of the American Statistical Association 55**: 789-98.

Gates CE and Bilbro JD (1978) Illustration of a cluster analysis method for mean separation. **Agronomy Journal 70**: 462-465.

Hotelling H (1933) Analysis of a complex of statistical variables into principal components. **Journal of Educational Psychology 24**: 417-441.

Jolliffe IT (1975) Cluster analysis as a multiple comparison method. **Applied Statistics, Proceedings of Conference at Dalhousie University 1**: 159-168.

Keuls M (1952) The use of the "studentized range" in connection with an analysis of variance. **Euphytica 1**: 112-122.

Leemis L and Trivedi KS (1996) A comparison of approximate interval estimators for the Bernoulli parameter. **The American Statistician Alexandria 50**: 63-68.

Metropolis N and Ulam S (1949) The Monte Carlo method. **Journal of the American Statistical Association 44**: 335-341.

Newman D (1939) The distribution of range in samples from a normal population expressed in terms of an independent estimate of standard deviation. **Biometrika 31**: 20-30.

O'Neill R and Wetherill GB (1971) The present state of multiple comparison methods. **Journal of the Royal Statistical Society 33**: 218-250.

Plackett RL (1971) The discussion on R O'Neill and G B Wetherill present state of multiple comparison methods. **Journal of the Royal Statistical Society 33**: 242-243.

Rao CR (1952) **Advanced statistical methods in biometric research**. John Wiley, New York, 390p.

Scheffé H (1953) A method for judging all contrasts in the analysis of variance. **Biometrika 40**: 87-110.

Santos C, Ferreira DF and Bueno-Filho JSS (2001) Novas alternativas de testes de agrupamento avaliadas por meio de simulação de Monte Carlo. **Ciência e Agrotecnologia 25**: 1382-1392.

SAS Institute (2011). SAS/IML 9.3 User's Guide. Sas Institute.

Scott AJ and Knott M (1974) A cluster analysis method for grouping means in the analysis of variance. **Biometrics 30**: 507-512.

Silva EC, Ferreira DF and Bearzotti E (1999) Avaliação do poder e taxas de erro tipo I do teste de Scott-Knott por meio do método de Monte Carlo. **Ciência e Agrotecnologia 23**: 687-696.

Student (1908) The probable error of a mean. **Biometrika 6**: 1-25.

Tasaki T, Yoden A and Goto M (1987) Graphical data analysis in comparative experimental studies. **Computational Statistics & Data Analysis 5**: 113-125.

Tukey JW (1949) Comparing individual means in the analysis of variance. **Biometrics 5**: 99-114.

Valdano SG and Di Rienzo J (2007) Discovering meaningful groups in hierarchical cluster analysis. An extension to the multivariate case of a multiple comparison method based on cluster analysis. **InterStat**: 1-28.

Willavise SA, Carmer SG and Walker WM (1980) Evaluation of cluster analysis for comparing treatment means. **Agronomy Journal 72**: 317-320.